Ridge & Lasso Regression

Presented by
Group 2,

Under Supervision of
Prof. Satya Prakash Singh

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Uttar Pradesh

February 11, 2024

Overview

Introduction and Motivation

- In OLS regression, the goal is to find the coefficients of the linear model that minimize the sum of the squared differences between the observed and the predicted values.
- However, when dealing with **high-dimensional data**, **dataset with many features**, **dataset with high multicollinearity**, there's a risk of overfitting. Regularization techniques are used to soften this problem. Regularization is a method used to add a penalty term to the objective function of the OLS regression.
- Ridge regression is a valuable tool for improving linear regression models' stability and generalization performance, making it a popular choice in data science and machine learning.
- Lasso Regression is commonly used in applications such as feature selection in machine learning, finance for portfolio optimization, and signal processing.

Ridge Regression (ℓ_2 regularization)

Objective function for ridge regression

$$Q(\beta) = (y - X\beta)^\intercal (y - X\beta) + \lambda \sum_{i=1}^{p} {\beta_i}^2.$$

- Ridge regression addresses the issue of multicollinearity by adding a penalty term to the ordinary least squares' objective function.
- This objective function of Ridge regression consists of two parts: the objective function of OLS, which measures the goodness of fit, and the penalty term, which is proportional to the sum of squared coefficients.

Ridge criterion

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \; Q(\beta). \tag{1}$$

- If $\hat{\beta}_{ridge}$ is the ridge regression estimator of β, then

$$\hat{\beta}_{ridge} = (X^\intercal X + \lambda \, \mathrm{I_p})^{-1} X^\intercal y. \tag{2}$$

The estimation of the ridge regression estimator depends upon the value of λ.

Properties of Ridge Regession

- $E(\hat{\beta}_{ridge}) = (X^\intercal X + \lambda\ \mathrm{I_p})^{-1} X^\intercal X\beta$. $\hat{\beta}_{ridge}$ is a biased estimator of β.
- $Var(\hat{\beta}_{ridge}) = \sigma^2 (X^\intercal X + \lambda\ \mathrm{I_p})^{-1} X^\intercal X (X^\intercal X + \lambda\ \mathrm{I_p})^{-1}$.
- Hoerl and Kennard [1] show that the total mean squared error for the ridge estimator is smaller than the corresponding least squared quantity.

Properties of Ridge Regression

- **Bias-Variance Tradeoff:** Ridge regression adds bias to the estimates but reduces variance, leading to better performance when multicollinearity exists.
- **Stabilization of Estimates:** Ridge regression stabilizes the estimates of regression coefficients by reducing their variance, especially when the predictors are highly correlated.
- **Shrinkage of Coefficients:** Ridge regression shrinks the coefficients towards zero by retaining all predictors in the model and preventing overfitting.

Lasso Regression (ℓ_1 regularization)

Objective function for Lasso Regression

$$Q(\beta) = (y - X\beta)^\intercal (y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|.$$

- This objective function consists of two parts: the objective function of OLS and the penalty term, which is proportional to the sum of the absolute values of the coefficients.

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} Q(\beta). \tag{3}$$

Properties of Lasso Regression

- **Variable Selection:** The penalty term encourages some coefficients to be exactly zero, effectively performing feature selection by shrinking less important coefficients towards zero.
- **Sparsity**: Due to its ability to set some coefficients to zero, Lasso regression often yields sparse models that are useful when dealing with high-dimensional datasets with many features, where automatic feature selection is desirable.
- **Flexibility in Model Complexity**: The choice of the regularization parameter λ controls the strength of the penalty and thus influences the balance between model simplicity and predictive accuracy.

Comparison

Parameter	**Ridge**	**Lasso**
Regularization Type	Uses ℓ_2 penalty.	Uses ℓ_1 penalty
Objective	To shrink the coefficients towards zero to reduce multicollinearity.	To shrink some coefficients towards zero for both variable reduction and model simplification.
Feature Selection	Does not perform feature selection: all features are included in the model, but their impact is minimized.	Performs feature selection: can completely eliminate some features by setting their coefficients to zero.
Suitability	Suitable in situations where all features are relevant, and there is multicollinearity.	Suitable when the number of predictors is high and there is a need to identify the most significant features.
Bias & Variance	Introduces bias but reduces variance.	Introduces bias but reduces variance, potentially more than Ridge due to feature elimination.
Model Complexity	Generally results in a more complex model compared to Lasso.	This leads to a simpler model, especially when irrelevant features are abundant.

Cross Validation

- To choose the tuning parameter (λ) in ridge regression, we introduce cross-validation.
- Cross-validation provides a way to estimate the test error without requiring new data.
- In cross-validation, our original dataset is broken into chunks to emulate independent datasets. Then the model is fit on some chunks and tested on other chunks, with the loss recorded. The way the data is broken into chunks can lead to different methods of cross-validation.

K-fold Cross Validation

- The data is randomly split into K roughly equal-sized parts.
- For any k^{th} split, the rest of the $K-1$ parts make up the training set, and the model fits the training set.
- We then estimate the prediction error for each element in the k^{th} part.
- Repeating this for all $k = 1, 2, \ldots, K$ parts, we have an estimate of the prediction error.

Real Data Application

- To demonstrate the application of ridge and lasso regression, we have used the weatherAUS dataset.
- Response variable: Rainfall Quantity.
  Predictor variables: Max Temp, Min Temp, Temperature, Windspeed, Humidity and Pressure at different time points of a day.
- We fit a multiple linear regression to the dataset containing 120381 rows and 12 columns.

```
> head(dat)
  MinTemp MaxTemp Rainfall WindGustSpeed WindSpeed9am WindSpeed3pm
1    13.4    22.9      0.6            44           20           24
2     7.4    25.1      0.0            44            4           22
3    12.9    25.7      0.0            46           19           26
4     9.2    28.0      0.0            24           11            9
5    17.5    32.3      1.0            41            7           20
6    14.6    29.7      0.2            56           19           24
  Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm
1          71          22      1007.7      1007.1    16.9    21.8
2          44          25      1010.6      1007.8    17.2    24.3
3          38          30      1007.6      1008.7    21.0    23.2
4          45          16      1017.6      1012.8    18.1    26.5
5          82          33      1010.8      1006.0    17.8    29.7
6          55          23      1009.2      1005.4    20.6    28.9
```

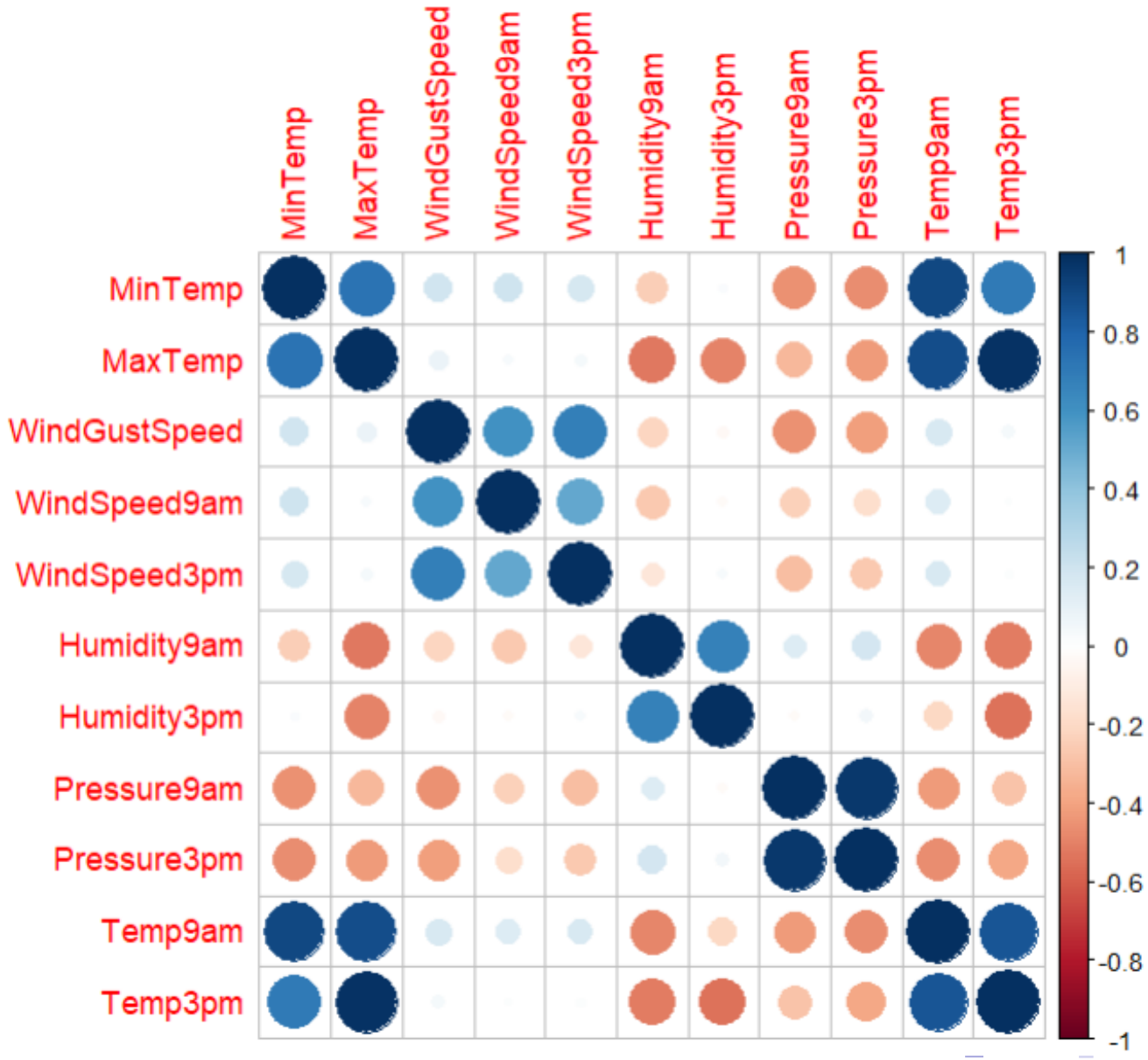
,

Multiple Linear Regression

```
> summary(fit_lm)

Call:
lm(formula = Rainfall ~ ., data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-13.80  -2.94  -1.04   0.94 353.69

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   109.980522   4.246712  25.898  < 2e-16 ***
MinTemp         0.015564   0.010695   1.455    0.146
MaxTemp        -0.586551   0.021597 -27.159  < 2e-16 ***
WindGustSpeed   0.100567   0.002791  36.038  < 2e-16 ***
WindSpeed9am    0.061318   0.003533  17.357  < 2e-16 ***
WindSpeed3pm   -0.098367   0.003763 -26.138  < 2e-16 ***
Humidity9am     0.095832   0.002414  39.704  < 2e-16 ***
Humidity3pm     0.061557   0.002746  22.417  < 2e-16 ***
Pressure9am    -0.405309   0.013861 -29.240  < 2e-16 ***
Pressure3pm     0.285719   0.013975  20.445  < 2e-16 ***
Temp9am         0.115180   0.016577   6.948 3.72e-12 ***
Temp3pm         0.619833   0.024161  25.655  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.883 on 120369 degrees of freedom
Multiple R-squared:  0.1346,    Adjusted R-squared:  0.1345
F-statistic:  1701 on 11 and 120369 DF,  p-value: < 2.2e-16
```

Multicollinearity

Correlation plot of the predictor variables:



Multicollinearity

- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A variance inflation factor (VIF) measures multicollinearity among the independent variables in a multiple regression model.
- $VIF_i = \frac{1}{1 - R_i^2}$, where R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

```
> vif(fit_lm)
      MinTemp       MaxTemp WindGustSpeed  WindSpeed9am
     8.972896     44.065294      2.744279      1.874924
 WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am
     2.066515     4.125090      6.253087     18.754508
  Pressure3pm       Temp9am       Temp3pm
    18.638572     22.115053     52.748113
```

Algorithm

Algorithm 1: Algorithm for choosing λ in ridge regression

Data: Data, Sequence of λ
Result: λ
for $1 \leq l \leq$ *Sequence of* λ **do**
  for $1 \leq i \leq 3$ **do**
    Divide the data into the training set and test set ;
    Calculate the beta estimate for the ridge regression ;
    Calculate the residual sum of squares ;
  end
  Calculate the average RSS ;
end
Choose the λ for which the average RSS is minimum

Ridge Estimates

- We consider the following sequence of λ for the cross-validation: $10^{-8}, \ldots, 10^{8}$.
- The value of λ is obtained as 0.01.
- Putting this value of λ, we get the value of $\hat{\beta}_{ridge}$ as :

```
> beta_ridge
                                  [,1]
rep(1, dim(dat)[1]) 109.66201219
MinTemp               0.01559528
MaxTemp              -0.58639915
WindGustSpeed         0.10064248
WindSpeed9am          0.06129837
WindSpeed3pm         -0.09838001
Humidity9am           0.09584285
Humidity3pm           0.06156381
Pressure9am          -0.40518802
Pressure3pm           0.28590405
Temp9am               0.11524120
Temp3pm               0.61974653
```

Lasso Estimates

- Since lasso regression does not provide estimates of β in closed form [2], hence we reach the estimates of β using the Newton Raphson method.
- We follow the same approach as in ridge regression using 3-fold cross-validation.
- We obtain the value of λ for which average RSS is minimum as : 0.01995262.
- $\hat{\beta}_{lasso}$:

```
> beta_lasso
                                [,1]
rep(1, dim(dat)[1]) 109.97472759
MinTemp               0.01556494
MaxTemp              -0.58654812
WindGustSpeed         0.10056794
WindSpeed9am          0.06131767
WindSpeed3pm         -0.09836769
Humidity9am           0.09583173
Humidity3pm           0.06155680
Pressure9am          -0.40530714
Pressure3pm           0.28572198
Temp9am               0.11518079
Temp3pm               0.61983174
```

Model Selection Criteria:

- $$\text{Mean Square Error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4}$$
- $$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{5}$$

Fitted model	MSE	MAE
Linear Regression	62.13368	3.468606
Ridge Regression	61.16653	3.451944
Lasso Regression	62.15027	3.459327

Table: Performance of MSE and MAE under different models

Conclusion

- There is high multicollinearity in our data.
- In this case, ridge regression gives us better estimates of the parameters.
- Lasso regression gives very close estimates of parameters as in linear regression.
- Among these three regression models, ridge regression gives the best performance in terms of both MSE and MAE.

References

Arthur E Hoerl and Robert W Kennard.
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67, 1970.

Robert Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288, 1996.

Thank You.